Chapter 1

An Overview of Regression Analysis

1.1 What Is Econometrics?

"Econometrics is too mathematical; it's the reason my best friend isn't majoring in economics."

"There are two things you are better off not watching in the making: sausages and econometric estimates."[1]

"Econometrics may be defined as the quantitative analysis of actual economic phenomena."[2]

"It's my experience that 'economy-tricks' is usually nothing more than a justification of what the author believed before the research was begun."

Obviously, econometrics means different things to different people. To beginning students, it may seem as if econometrics is an overly complex obstacle to an otherwise useful education. To skeptical observers, econometric

1. Ed Leamer, "Let's take the Con out of Econometrics," *American Economic Review*, Vol. 73, No. 1, p. 37.

2. Paul A. Samuelson, T. C. Koopmans, and J. R. Stone, "Report of the Evaluative Committee for *Econometrica*," *Econometrica*, 1954, p. 141.

results should be trusted only when the steps that produced those results are completely known. To professionals in the field, econometrics is a fascinating set of techniques that allows the measurement and analysis of economic phenomena and the prediction of future economic trends.

You're probably thinking that such diverse points of view sound like the statements of blind people trying to describe an elephant based on which part they happen to be touching, and you're partially right. Econometrics has both a formal definition and a larger context. Although you can easily memorize the formal definition, you'll get the complete picture only by understanding the many uses of and alternative approaches to econometrics.

That said, we need a formal definition. **Econometrics**—literally, "economic measurement"—is the quantitative measurement and analysis of actual economic and business phenomena. It attempts to quantify economic reality and bridge the gap between the abstract world of economic theory and the real world of human activity. To many students, these worlds may seem far apart. On the one hand, economists theorize equilibrium prices based on carefully conceived marginal costs and marginal revenues; on the other, many firms seem to operate as though they have never heard of such concepts. Econometrics allows us to examine data and to quantify the actions of firms, consumers, and governments. Such measurements have a number of different uses, and an examination of these uses is the first step to understanding econometrics.

Uses of Econometrics

Econometrics has three major uses:

1. describing economic reality
2. testing hypotheses about economic theory and policy
3. forecasting future economic activity

The simplest use of econometrics is description. We can use econometrics to quantify economic activity and measure marginal effects because econometrics allows us to estimate numbers and put them in equations that previously contained only abstract symbols. For example, consumer demand for a particular product often can be thought of as a relationship between the quantity demanded (Q) and the product's price (P), the price of a substitute (P_s), and disposable income (Yd). For most goods, the relationship between consumption and disposable income is expected to be positive, because an increase in disposable income will be associated with an increase in the consumption of the product. Econometrics actually allows us to estimate that

relationship based upon past consumption, income, and prices. In other words, a general and purely theoretical functional relationship like:

$$Q = \beta_0 + \beta_1 P + \beta_2 P_S + \beta_1 Yd \tag{1.1}$$

can become explicit:

$$Q = 27.7 - 0.11P + 0.03P_S + 0.23Yd \tag{1.2}$$

This technique gives a much more specific and descriptive picture of the function.[3] Let's compare Equations 1.1 and 1.2. Instead of expecting consumption merely to "increase" if there is an increase in disposable income, Equation 1.2 allows us to expect an increase of a specific amount (0.23 units for each unit of increased disposable income). The number 0.23 is called an estimated regression coefficient, and it is the ability to estimate these coefficients that makes econometrics valuable.

The second use of econometrics is hypothesis testing, the evaluation of alternative theories with quantitative evidence. Much of economics involves building theoretical models and testing them against evidence, and hypothesis testing is vital to that scientific approach. For example, you could test the hypothesis that the product in Equation 1.1 is what economists call a normal good (one for which the quantity demanded increases when disposable income increases). You could do this by applying various statistical tests to the estimated coefficient (0.23) of disposable income (Yd) in Equation 1.2. At first glance, the evidence would seem to support this hypothesis, because the coefficient's sign is positive, but the "statistical significance" of that estimate would have to be investigated before such a conclusion could be justified. Even though the estimated coefficient is positive, as expected, it may not be sufficiently different from zero to convince us that the true coefficient is indeed positive.

The third and most difficult use of econometrics is to forecast or predict what is likely to happen next quarter, next year, or further into the future, based on what has happened in the past. For example, economists use econometric models to make forecasts of variables like sales, profits, Gross Domestic Product (GDP), and the inflation rate. The accuracy of such forecasts depends in large measure on the degree to which the past is a good guide to the future. Business leaders and politicians tend to be especially interested in this use of

3. It's of course naïve to build a model of sales (demand) without taking supply into consideration. Unfortunately, it's very difficult to learn how to estimate a system of simultaneous equations until you've learned how to estimate a single equation. As a result, we will postpone our discussion of the econometrics of simultaneous equations until Chapter 14. Until then, you should be aware that we sometimes will encounter right-hand-side variables that are not truly "independent" from a theoretical point of view.

econometrics because they need to make decisions about the future, and the penalty for being wrong (bankruptcy for the entrepreneur and political defeat for the candidate) is high. To the extent that econometrics can shed light on the impact of their policies, business and government leaders will be better equipped to make decisions. For example, if the president of a company that sold the product modeled in Equation 1.1 wanted to decide whether to increase prices, forecasts of sales with and without the price increase could be calculated and compared to help make such a decision.

Alternative Econometric Approaches

There are many different approaches to quantitative work. For example, the fields of biology, psychology, and physics all face quantitative questions similar to those faced in economics and business. However, these fields tend to use somewhat different techniques for analysis because the problems they face aren't the same. For example, economics typically is an observational discipline rather than an experimental one. "We need a special field called econometrics, and textbooks about it, because it is generally accepted that economic data possess certain properties that are not considered in standard statistics texts or are not sufficiently emphasized there for use by economists."[4]

Different approaches also make sense within the field of economics. A model built solely for descriptive purposes might be different from a forecasting model, for example.

To get a better picture of these approaches, let's look at the steps used in nonexperimental quantitative research:

1. specifying the models or relationships to be studied
2. collecting the data needed to quantify the models
3. quantifying the models with the data

The specifications used in step 1 and the techniques used in step 3 differ widely between and within disciplines. Choosing the best specification for a given model is a theory-based skill that is often referred to as the "art" of econometrics. There are many alternative approaches to quantifying the same equation, and each approach may produce somewhat different results. The choice of approach is left to the individual econometrician (the researcher using econometrics), but each researcher should be able to justify that choice.

4. Clive Granger, "A Review of Some Recent Textbooks of Econometrics," *Journal of Economic Literature*, Vol. 32, No. 1, p. 117.

This book will focus primarily on one particular econometric approach: *single-equation linear regression analysis.* The majority of this book will thus concentrate on regression analysis, but it is important for every econometrician to remember that regression is only one of many approaches to econometric quantification.

The importance of critical evaluation cannot be stressed enough; a good econometrician can diagnose faults in a particular approach and figure out how to repair them. The limitations of the regression analysis approach must be fully perceived and appreciated by anyone attempting to use regression analysis or its findings. The possibility of missing or inaccurate data, incorrectly formulated relationships, poorly chosen estimating techniques, or improper statistical testing procedures implies that the results from regression analyses always should be viewed with some caution.

1.2 What Is Regression Analysis?

Econometricians use regression analysis to make quantitative estimates of economic relationships that previously have been completely theoretical in nature. After all, anybody can claim that the quantity of iPhones demanded will increase if the price of those phones decreases (holding everything else constant), but not many people can put specific numbers into an equation and estimate *by how many* iPhones the quantity demanded will increase for each dollar that price decreases. To predict the *direction* of the change, you need a knowledge of economic theory and the general characteristics of the product in question. To predict the *amount* of the change, though, you need a sample of data, and you need a way to estimate the relationship. The most frequently used method to estimate such a relationship in econometrics is regression analysis.

Dependent Variables, Independent Variables, and Causality

Regression analysis is a statistical technique that attempts to "explain" movements in one variable, the **dependent variable**, as a function of movements in a set of other variables, called the **independent** (or **explanatory**) **variables**, through the quantification of one or more equations. For example, in Equation 1.1:

$$Q = \beta_0 + \beta_1 P + \beta_2 P_S + \beta_1 Yd \qquad (1.1)$$

Q is the dependent variable and P, P_S, and Yd are the independent variables. Regression analysis is a natural tool for economists because most (though not all) economic propositions can be stated in such equations. For example, the quantity demanded (dependent variable) is a function of price, the prices of substitutes, and income (independent variables).

Much of economics and business is concerned with cause-and-effect propositions. If the price of a good increases by one unit, then the quantity demanded decreases on average by a certain amount, depending on the price elasticity of demand (defined as the percentage change in the quantity demanded that is caused by a one percent increase in price). Similarly, if the quantity of capital employed increases by one unit, then output increases by a certain amount, called the marginal productivity of capital. Propositions such as these pose an if-then, or causal, relationship that logically postulates that a dependent variable's movements are determined by movements in a number of specific independent variables.

Don't be deceived by the words "dependent" and "independent," however. Although many economic relationships are causal by their very nature, a regression result, no matter how statistically significant, cannot prove causality. All regression analysis can do is test whether a significant quantitative relationship exists. Judgments as to causality must also include a healthy dose of economic theory and common sense. For example, the fact that the bell on the door of a flower shop rings just before a customer enters and purchases some flowers by no means implies that the bell causes purchases! If events A and B are related statistically, it may be that A causes B, that B causes A, that some omitted factor causes both, or that a chance correlation exists between the two.

The cause-and-effect relationship often is so subtle that it fools even the most prominent economists. For example, in the late nineteenth century, English economist Stanley Jevons hypothesized that sunspots caused an increase in economic activity. To test this theory, he collected data on national output (the dependent variable) and sunspot activity (the independent variable) and showed that a significant positive relationship existed. This result led him, and some others, to jump to the conclusion that sunspots did indeed cause output to rise. Such a conclusion was unjustified because regression analysis cannot confirm causality; it can only test the strength and direction of the quantitative relationships involved.

Single-Equation Linear Models

The simplest single-equation regression model is:

$$Y = \beta_0 + \beta_1 X \tag{1.3}$$

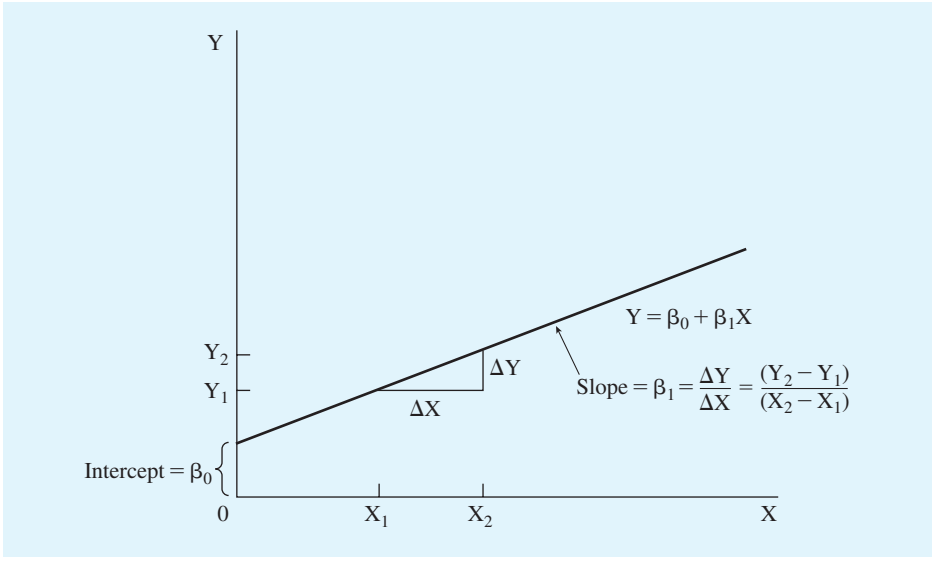


Figure 1.1 Graphical Representation of the Coefficients of the Regression Line

The graph of the equation $Y = \beta_0 + \beta_1 X$ is linear with a constant slope equal to $\beta_1 = \Delta Y / \Delta X$.

Equation 1.3 states that Y, the dependent variable, is a single-equation linear function of X, the independent variable. The model is a single-equation model because it's the only equation specified. The model is linear because if you were to plot Equation 1.3 it would be a straight line rather than a curve.

The βs are the coefficients that determine the coordinates of the straight line at any point. β_0 is the **constant** or **intercept** term; it indicates the value of Y when X equals zero. β_1 is the **slope coefficient**, and it indicates the amount that Y will change when X increases by one unit. The line in Figure 1.1 illustrates the relationship between the coefficients and the graphical meaning of the regression equation. As can be seen from the diagram, Equation 1.3 is indeed linear.

The slope coefficient, β_1, shows the response of Y to a one-unit increase in X. Much of the emphasis in regression analysis is on slope coefficients such as β_1. In Figure 1.1 for example, if X were to increase by one from X_1 to X_2 (ΔX), the value of Y in Equation 1.3 would increase from Y_1 to Y_2 (ΔY). For linear (i.e., straight-line) regression models, the response in the predicted value of Y due to a change in X is constant and equal to the slope coefficient β_1:

$$\frac{(Y_2 - Y_1)}{(X_2 - X_1)} = \frac{\Delta Y}{\Delta X} = \beta_1$$

where Δ is used to denote a change in the variables. Some readers may recognize this as the "rise" (ΔY) divided by the "run" (ΔX). For a linear model, the slope is constant over the entire function.

If linear regression techniques are going to be applied to an equation, that equation *must* be linear. An equation is **linear** if plotting the function in terms of X and Y generates a straight line; for example, Equation 1.3 is linear.[5]

$$Y = \beta_0 + \beta_1 X \tag{1.3}$$

The Stochastic Error Term

Besides the variation in the dependent variable (Y) that is caused by the independent variable (X), there is almost always variation that comes from other sources as well. This additional variation comes in part from omitted explanatory variables (e.g., X_2 and X_3). However, even if these extra variables are added to the equation, there still is going to be some variation in Y that simply cannot be explained by the model.[6] This variation probably comes from sources such as omitted influences, measurement error, incorrect functional form, or purely random and totally unpredictable occurrences. By *random* we mean something that has its value determined entirely by chance.

Econometricians admit the existence of such inherent unexplained variation ("error") by explicitly including a stochastic (or random) error term in their regression models. A **stochastic error term** is a term that is added to a regression equation to introduce all of the variation in Y that cannot be explained by the included Xs. It is, in effect, a symbol of the econometrician's ignorance or inability to model all the movements of the dependent variable. The error term (sometimes called a disturbance term) usually is referred to with the symbol epsilon (ϵ), although other symbols (like u or v) sometimes are used.

---

5. Technically, as you will learn in Chapter 7, this equation is linear in the coefficients β_0 and β_1 and linear in the variables Y and X. The application of regression analysis to equations that are nonlinear in the variables is covered in Chapter 7. The application of regression techniques to equations that are nonlinear in the coefficients, however, is much more difficult.

6. The exception would be the extremely rare case where the data can be explained by some sort of physical law and are measured perfectly. Here, continued variation would point to an omitted independent variable. A similar kind of problem is often encountered in astronomy, where planets can be discovered by noting that the orbits of known planets exhibit variations that can be caused only by the gravitational pull of another heavenly body. Absent these kinds of physical laws, researchers in economics and business would be foolhardy to believe that *all* variation in Y can be explained by a regression model because there are always elements of error in any attempt to measure a behavioral relationship.

The addition of a stochastic error term (ϵ) to Equation 1.3 results in a typical regression equation:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1.4}$$

Equation 1.4 can be thought of as having two components, the *deterministic* component and the *stochastic*, or random, component. The expression $\beta_0 + \beta_1 X$ is called the *deterministic* component of the regression equation because it indicates the value of Y that is determined by a given value of X, which is assumed to be nonstochastic. This deterministic component can also be thought of as the **expected value** of Y given X, the mean value of the Ys associated with a particular value of X. For example, if the average height of all 13-year-old girls is 5 feet, then 5 feet is the expected value of a girl's height given that she is 13. The deterministic part of the equation may be written:

$$E(Y|X) = \beta_0 + \beta_1 X \tag{1.5}$$

which states that the expected value of Y given X, denoted as $E(Y|X)$, is a linear function of the independent variable (or variables if there are more than one).

Unfortunately, the value of Y observed in the real world is unlikely to be exactly equal to the deterministic expected value $E(Y|X)$. After all, not all 13-year-old girls are 5 feet tall. As a result, the stochastic element (ϵ) must be added to the equation:

$$Y = E(Y|X) + \epsilon = \beta_0 + \beta_1 X + \epsilon \tag{1.6}$$

The stochastic error term must be present in a regression equation because there are at least four sources of variation in Y other than the variation in the included Xs:

1. Many minor influences on Y are *omitted* from the equation (for example, because data are unavailable).
2. It is virtually impossible to avoid some sort of *measurement error* in the dependent variable.
3. The underlying theoretical equation might have a *different functional form* (or shape) than the one chosen for the regression. For example, the underlying equation might be nonlinear.
4. All attempts to generalize human behavior must contain at least some amount of unpredictable or *purely random* variation.

To get a better feeling for these components of the stochastic error term, let's think about a consumption function (aggregate consumption as a function of aggregate disposable income). First, consumption in a particular year may have been less than it would have been because of uncertainty over the future course of the economy. Since this uncertainty is hard to measure, there might be no variable measuring consumer uncertainty in the equation. In such a case, the impact of the omitted variable (consumer uncertainty) would likely end up in the stochastic error term. Second, the observed amount of consumption may have been different from the actual level of consumption in a particular year due to an error (such as a sampling error) in the measurement of consumption in the National Income Accounts. Third, the underlying consumption function may be nonlinear, but a linear consumption function might be estimated. (To see how this incorrect functional form would cause errors, see Figure 1.2.) Fourth, the consumption function

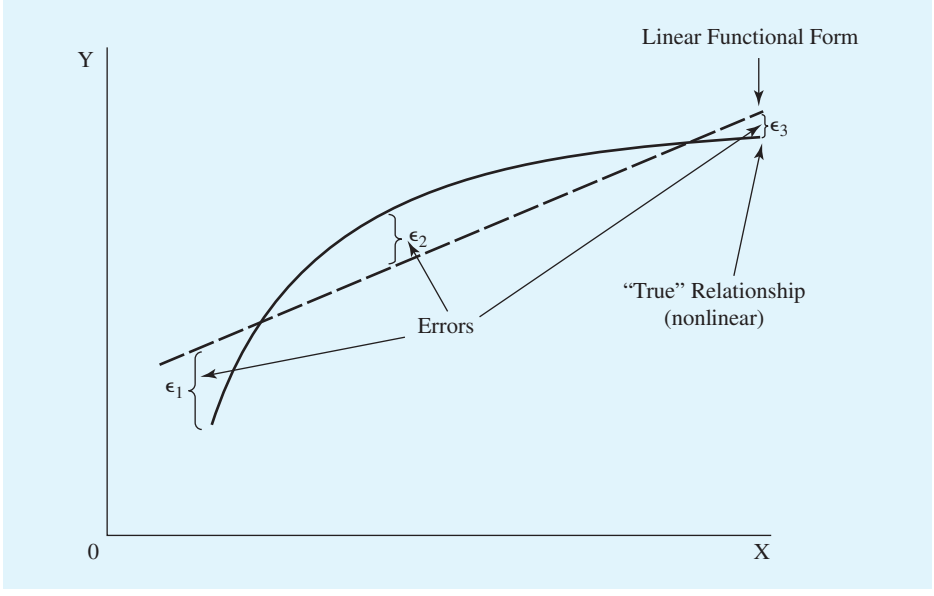


Figure 1.2 Errors Caused by Using a Linear Functional Form to Model a Nonlinear Relationship

One source of stochastic error is the use of an incorrect functional form. For example, if a linear functional form is used when the underlying relationship is nonlinear, systematic errors (the ϵs) will occur. These nonlinearities are just one component of the stochastic error term. The others are omitted variables, measurement error, and purely random variation.

attempts to portray the behavior of people, and there is always an element of unpredictability in human behavior. At any given time, some random event might increase or decrease aggregate consumption in a way that might never be repeated and couldn't be anticipated.

These possibilities explain the existence of a difference between the observed values of Y and the values expected from the deterministic component of the equation, $E(Y|X)$. These sources of error will be covered in more detail in the following chapters, but for now it is enough to recognize that in econometric research there will always be some stochastic or random element, and, for this reason, an error term must be added to all regression equations.

Extending the Notation

Our regression notation needs to be extended to allow the possibility of more than one independent variable and to include reference to the number of observations. A typical observation (or unit of analysis) is an individual person, year, or country. For example, a series of annual observations starting in 1985 would have $Y_1 = Y$ for 1985, Y_2 for 1986, etc. If we include a specific reference to the observations, the single-equation linear regression model may be written as:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad (i = 1, 2, \ldots, N) \tag{1.7}$$

where:
- Y_i = the *i*th observation of the dependent variable
- X_i = the *i*th observation of the independent variable
- ϵ_i = the *i*th observation of the stochastic error term
- β_0, β_1 = the regression coefficients
- N = the number of observations

This equation is actually N equations, one for each of the N observations:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \epsilon_2 \\ Y_3 &= \beta_0 + \beta_1 X_3 + \epsilon_3 \\ &\vdots \\ Y_N &= \beta_0 + \beta_1 X_N + \epsilon_N \end{aligned}$$

That is, the regression model is assumed to hold for each observation. The coefficients do not change from observation to observation, but the values of Y, X, and ϵ do.

A second notational addition allows for more than one independent variable. Since more than one independent variable is likely to have an effect on the dependent variable, our notation should allow these additional explanatory Xs to be added. If we define:

X_{1i} = the ith observation of the first independent variable
X_{2i} = the ith observation of the second independent variable
X_{3i} = the ith observation of the third independent variable

then all three variables can be expressed as determinants of Y.

The resulting equation is called a **multivariate** (more than one independent variable) linear **regression model:**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i \tag{1.8}$$

The *meaning of the regression coefficient* β_1 in this equation is the impact of a one-unit increase in X_1 on the dependent variable Y, *holding constant* X_2 and X_3. Similarly, β_2 gives the impact of a one-unit increase in X_2 on Y, holding X_1 and X_3 constant.

These *multivariate regression coefficients* (which are parallel in nature to partial derivatives in calculus) serve to isolate the impact on Y of a change in one variable from the impact on Y of changes in the other variables. This is possible because multivariate regression takes the movements of X_2 and X_3 into account when it estimates the coefficient of X_1. The result is quite similar to what we would obtain if we were capable of conducting controlled laboratory experiments in which only one variable at a time was changed.

In the real world, though, it is very difficult to run controlled economic experiments,[7] because many economic factors change simultaneously, often in opposite directions. Thus the ability of regression analysis to measure the impact of one variable on the dependent variable, *holding constant the influence of the other variables in the equation,* is a tremendous advantage. Note that if a variable is not included in an equation, then its impact is *not* held constant in the estimation of the regression coefficients. This will be discussed further in Chapter 6.

7. Such experiments are difficult but not impossible. See Section 16.1.

This material is pretty abstract, so let's look at two examples. As a first example, consider an equation with only one independent variable, a model of a person's weight as a function of their height. The theory behind this equation is that, other things being equal, the taller a person is the more they tend to weigh.

The dependent variable in such an equation would be the weight of the person, while the independent variable would be that person's height:

$$\text{Weight}_i = \beta_0 + \beta_1 \text{Height}_i + \epsilon_i \tag{1.9}$$

What exactly do the "i" subscripts mean in Equation 1.9? Each value of i refers to a different person in the sample, so another way to think about the subscripts is that:

$$\begin{aligned}
\text{Weight}_{\text{woody}} &= \beta_0 + \beta_1 \text{Height}_{\text{woody}} + \epsilon_{\text{woody}} \\
\text{Weight}_{\text{lesley}} &= \beta_0 + \beta_1 \text{Height}_{\text{lesley}} + \epsilon_{\text{lesley}} \\
\text{Weight}_{\text{bruce}} &= \beta_0 + \beta_1 \text{Height}_{\text{bruce}} + \epsilon_{\text{bruce}} \\
\text{Weight}_{\text{mary}} &= \beta_0 + \beta_1 \text{Height}_{\text{mary}} + \epsilon_{\text{mary}}
\end{aligned}$$

Take a look at these equations. Each person (observation) in the sample has their own individual weight and height; that makes sense. But why does each person have their own value for ϵ, the stochastic error term? The answer is that random events (like those expressed by ϵ) impact people differently, so each person needs to have their own value of ϵ in order to reflect these differences. In contrast, note that the subscripts of the regression coefficients (the βs) don't change from person to person but instead apply to the entire sample. We'll learn more about this equation in Section 1.4.

As a second example, let's look at an equation with more than one independent variable. Suppose we want to understand how wages are determined in a particular field, perhaps because we think that there might be discrimination in that field. The wage of a worker would be the dependent variable (WAGE), but what would be good independent variables? What variables would influence a person's wage in a given field? Well, there are literally dozens of reasonable possibilities, but three of the most common are the work experience (EXP), education (EDU), and gender (GEND) of the worker, so let's use these. To create a regression equation with these variables, we'd redefine the variables in Equation 1.8 to meet our definitions:

Y = WAGE = the wage of the worker
X_1 = EXP = the years of work experience of the worker
X_2 = EDU = the years of education beyond high school of the worker
X_3 = GEND = the gender of the worker (1 = male and 0 = female)

The last variable, GEND, is unusual in that it can take on only two values, 0 and 1; this kind of variable is called a dummy variable, and it's extremely useful when we want to quantify a concept that is inherently qualitative (like gender). We'll discuss dummy variables in more depth in Sections 3.3 and 7.4.

If we substitute these definitions into Equation 1.8, we get:

$$WAGE_i = \beta_0 + \beta_1 EXP_i + \beta_2 EDU_i + \beta_3 GEND_i + \epsilon_i \qquad (1.10)$$

Equation 1.10 specifies that a worker's wage is a function of the experience, education, and gender of that worker. In such an equation, what would the meaning of β_1 be? Some readers will guess that β_1 measures the amount by which the average wage increases for an additional year of experience, but such a guess would miss the fact that there are two other independent variables in the equation that also explain wages. The correct answer is that β_1 gives us the impact on wages of a one-year increase in experience, *holding constant* education and gender. This is a significant difference, because it allows researchers to control for specific complicating factors without running controlled experiments.

Before we conclude this section, it's worth noting that the general multivariate regression model with K independent variables is written as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \epsilon_i \qquad (1.11)$$

where i goes from 1 to N and indicates the observation number.

If the sample consists of a series of years or months (called a time series), then the subscript i is usually replaced with a t to denote time.[8]

1.3 The Estimated Regression Equation

Once a specific equation has been decided upon, it must be quantified. This quantified version of the theoretical regression equation is called the **estimated regression equation** and is obtained from a sample of data for actual Xs and Ys. Although the theoretical equation is purely abstract in nature:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad (1.12)$$

8. The order of the subscripts doesn't matter as long as the appropriate definitions are presented. We prefer to list the variable number first (X_{1i}) because we think it's easier for a beginning econometrician to understand. However, as the reader moves on to matrix algebra and computer spreadsheets, it will become common to list the observation number first, as in X_{i1}. Often the observational subscript is deleted, and the reader is expected to understand that the equation holds for each observation in the sample.

the estimated regression equation has actual numbers in it:

$$\hat{Y}_i = 103.40 + 6.38X_i \tag{1.13}$$

The observed, real-world values of X and Y are used to calculate the coefficient estimates 103.40 and 6.38. These estimates are used to determine $\hat{Y}$ (read as "Y-hat"), the *estimated* or *fitted* value of Y.

Let's look at the differences between a theoretical regression equation and an estimated regression equation. First, the theoretical regression coefficients β_0 and β_1 in Equation 1.12 have been replaced with *estimates* of those coefficients like 103.40 and 6.38 in Equation 1.13. We can't actually observe the values of the true[9] regression coefficients, so instead we calculate estimates of those coefficients from the data. The estimated regression coefficients, more generally denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$ (read as "beta-hats"), are empirical best guesses of the true regression coefficients and are obtained from data from a sample of the Ys and Xs. The expression

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{1.14}$$

is the empirical counterpart of the theoretical regression Equation 1.12. The calculated estimates in Equation 1.13 are examples of the estimated regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$. For each sample we calculate a different set of estimated regression coefficients.

$\hat{Y}_i$ is the *estimated value* of Y_i, and it represents the value of Y calculated from the estimated regression equation for the *i*th observation. As such, $\hat{Y}_i$ is our prediction of $E(Y_i|X_i)$ from the regression equation. The closer these $\hat{Y}$s are to the Ys in the sample, the better the fit of the equation. (The word *fit* is used here much as it would be used to describe how well clothes fit.)

The difference between the estimated value of the dependent variable $(\hat{Y}_i)$ and the actual value of the dependent variable (Y_i) is defined as the **residual** (e_i):

$$e_i = Y_i - \hat{Y}_i \tag{1.15}$$

9. Our use of the word "true" throughout the text should be taken with a grain of salt. Many philosophers argue that the concept of truth is useful only relative to the scientific research program in question. Many economists agree, pointing out that what is true for one generation may well be false for another. To us, the true coefficient is the one that you'd obtain if you could run a regression on the entire relevant population. Thus, readers who so desire can substitute the phrase "population coefficient" for "true coefficient" with no loss in meaning.

Note the distinction between the residual in Equation 1.15 and the error term:

$$\epsilon_i = Y_i - E(Y_i|X_i) \tag{1.16}$$

The *residual* is the difference between the observed Y and the estimated regression line $(\hat{Y})$, while the *error term* is the difference between the observed Y and the true regression equation (the expected value of Y). Note that the error term is a theoretical concept that can never be observed, but the residual is a real-world value that is calculated for each observation every time a regression is run. The residual can be thought of as an estimate of the error term, and e could have been denoted as $\hat{\epsilon}$. Most regression techniques not only calculate the residuals but also attempt to compute values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that keep the residuals as low as possible. The smaller the residuals, the better the fit, and the closer the $\hat{Y}$s will be to the Ys.

All these concepts are shown in Figure 1.3. The (X, Y) pairs are shown as points on the diagram, and both the true regression equation (which

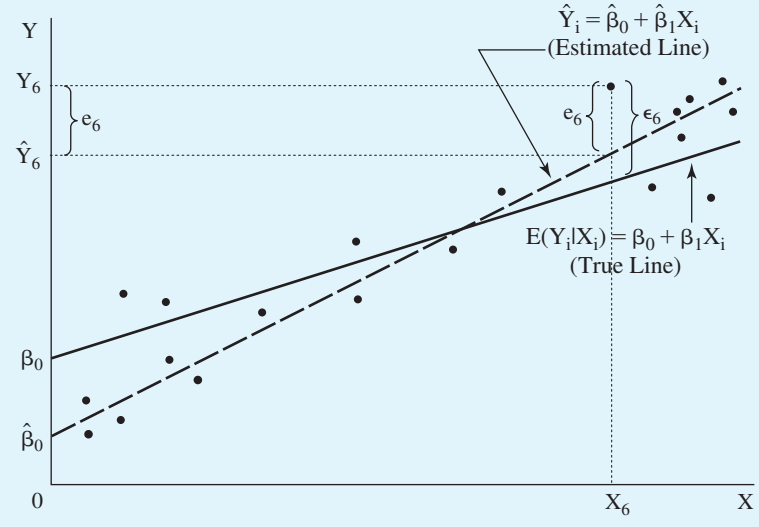


Figure 1.3 True and Estimated Regression Lines

The true relationship between X and Y (the solid line) typically cannot be observed, but the estimated regression line (the dashed line) can. The difference between an observed data point (for example, i = 6) and the true line is the value of the stochastic error term (ϵ_6). The difference between the observed Y_6 and the estimated value from the regression line $(\hat{Y}_6)$ is the value of the residual for this observation, e_6.

cannot be seen in real applications) and an estimated regression equation are included. Notice that the estimated equation is close to but not equivalent to the true line. This is a typical result.

In Figure 1.3, $\hat{Y}_6$, the computed value of Y for the sixth observation, lies on the estimated (dashed) line, and it differs from Y_6, the actual observed value of Y for the sixth observation. The difference between the observed and estimated values is the residual, denoted by e_6. In addition, although we usually would not be able to see an observation of the error term, we have drawn the assumed true regression line here (the solid line) to see the sixth observation of the error term, ϵ_6, which is the difference between the true line and the observed value of Y, Y_6.

The following table summarizes the notation used in the true and estimated regression equations:

True Regression Equation	Estimated Regression Equation
β_0	$\hat{\beta}_0$
β_1	$\hat{\beta}_1$
ϵ_i	e_i

The estimated regression model can be extended to more than one independent variable by adding the additional Xs to the right side of the equation. The multivariate estimated regression counterpart of Equation 1.14 is:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_K X_{Ki} \qquad (1.17)$$

Diagrams of such multivariate equations, by the way, are not possible for more than two independent variables and are quite awkward for exactly two independent variables.

1.4 A Simple Example of Regression Analysis

Let's look at a fairly simple example of regression analysis. Suppose you've accepted a summer job as a weight guesser at the local amusement park, Magic Hill. Customers pay two dollars each, which you get to keep if you guess their weight within 10 pounds. If you miss by more than 10 pounds, then you have to return the two dollars and give the customer a small prize that you buy from Magic Hill for three dollars each. Luckily, the friendly managers of Magic Hill have arranged a number of marks on the wall behind the customer so that you are capable of measuring the customer's

height accurately. Unfortunately, there is a five-foot wall between you and the customer, so you can tell little about the person except for height and (usually) gender.

On your first day on the job, you do so poorly that you work all day and somehow manage to lose two dollars, so on the second day you decide to collect data to run a regression to estimate the relationship between weight and height. Since most of the participants are male, you decide to limit your sample to males. You hypothesize the following theoretical relationship:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad (1.18)$$

where: Y_i = the weight (in pounds) of the *i*th customer
X_i = the height (in inches above 5 feet) of the *i*th customer
ϵ_i = the value of the stochastic error term for the *i*th customer

In this case, the sign of the theoretical relationship between height and weight is believed to be positive (signified by the positive sign above β_1 in the general theoretical equation), but you must quantify that relationship in order to estimate weights when given heights. To do this, you need to collect a data set, and you need to apply regression analysis to the data.

The next day you collect the data summarized in Table 1.1 and run your regression on the Magic Hill computer, obtaining the following estimates:

$$\hat{\beta}_0 = 103.40 \qquad \hat{\beta}_1 = 6.38$$

This means that the equation

$$\text{Estimated weight} = 103.40 + 6.38 \cdot \text{Height (inches above five feet)} \qquad (1.19)$$

is worth trying as an alternative to just guessing the weights of your customers. Such an equation estimates weight with a constant base of 103.40 pounds and adds 6.38 pounds for every inch of height over 5 feet. Note that the sign of $\hat{\beta}_1$ is positive, as you expected.

How well does the equation work? To answer this question, you need to calculate the residuals (Y_i minus $\hat{Y}_i$) from Equation 1.19 to see how many were greater than ten. As can be seen in the last column in Table 1.1, if you had applied the equation to these 20 people, you wouldn't exactly have gotten rich, but at least you would have earned $25.00 instead of losing $2.00. Figure 1.4 shows not only Equation 1.19 but also the weight and height data for all 20 customers used as the sample. With a different group of people, the results would of course be different.

Equation 1.19 would probably help a beginning weight guesser, but it could be improved by adding other variables or by collecting a larger sample.

Table 1.1 Data for and Results of the Weight-Guessing Equation

Observation i (1)	Height Above 5′ X_i (2)	Weight Y_i (3)	Predicted Weight $\hat{Y}_i$ (4)	Residual e_i (5)	$ Gain or Loss (6)
1	5.0	140.0	135.3	4.7	+2.00
2	9.0	157.0	160.8	−3.8	+2.00
3	13.0	205.0	186.3	18.7	−3.00
4	12.0	198.0	179.9	18.1	−3.00
5	10.0	162.0	167.2	−5.2	+2.00
6	11.0	174.0	173.6	0.4	+2.00
7	8.0	150.0	154.4	−4.4	+2.00
8	9.0	165.0	160.8	4.2	+2.00
9	10.0	170.0	167.2	2.8	+2.00
10	12.0	180.0	179.9	0.1	+2.00
11	11.0	170.0	173.6	−3.6	+2.00
12	9.0	162.0	160.8	1.2	+2.00
13	10.0	165.0	167.2	−2.2	+2.00
14	12.0	180.0	179.9	0.1	+2.00
15	8.0	160.0	154.4	5.6	+2.00
16	9.0	155.0	160.8	−5.8	+2.00
17	10.0	165.0	167.2	−2.2	+2.00
18	15.0	190.0	199.1	−9.1	+2.00
19	13.0	185.0	186.3	−1.3	+2.00
20	11.0	155.0	173.6	−18.6	−3.00
				TOTAL =	$25.00

Note: This data set, and every other data set in the text, is available on the text's website in four formats. Datafile = HTWT1

Such an equation is realistic, though, because it's likely that every successful weight guesser uses an equation like this without consciously thinking about that concept.

Our goal with this equation was to quantify the theoretical weight/height equation, Equation 1.18, by collecting data (Table 1.1) and calculating an estimated regression, Equation 1.19. Although the true equation, like observations of the stochastic error term, can never be known, we were able to come up with an estimated equation that had the sign we expected for $\hat{\beta}_1$ and that helped us in our job. Before you decide to quit school or your job and try to make your living guessing weights at Magic Hill, there is quite a bit more to learn about regression analysis, so we'd better move on.

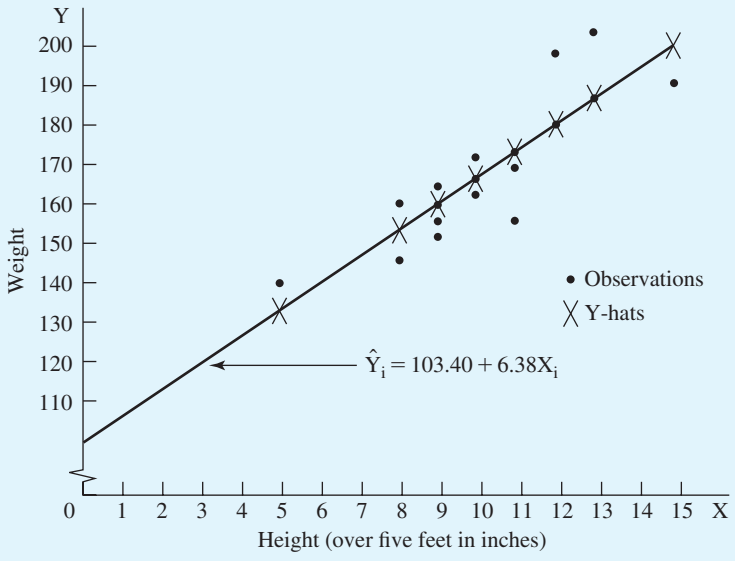


Figure 1.4 A Weight-Guessing Equation

If we plot the data from the weight-guessing example and include the estimated regression line, we can see that the estimated $\hat{Y}$s come fairly close to the observed Ys for all but three observations. Find a male friend's height and weight on the graph. How well does the regression equation work?

1.5 Using Regression to Explain Housing Prices

As much fun as guessing weights at an amusement park might be, it's hardly a typical example of the use of regression analysis. For every regression run on such an off-the-wall topic, there are literally hundreds run to *describe* the reaction of GDP to an increase in the money supply, to *test* an economic theory with new data, or to *forecast* the effect of a price change on a firm's sales.

As a more realistic example, let's look at a model of housing prices. The purchase of a house is probably the most important financial decision in an individual's life, and one of the key elements in that decision is an appraisal of the house's value. If you overvalue the house, you can lose thousands of dollars by paying too much; if you undervalue the house, someone might outbid you.

All this wouldn't be much of a problem if houses were homogeneous products, like corn or gold, that have generally known market prices with which to compare a particular asking price. Such is hardly the case in the real estate market. Consequently, an important element of every housing

purchase is an appraisal of the market value of the house, and many real estate appraisers use regression analysis to help them in their work.

Suppose your family is about to buy a house, but you're convinced that the owner is asking too much money. The owner says that the asking price of $230,000 is fair because a larger house next door sold for $230,000 about a year ago. You're not sure it's reasonable to compare the prices of different-sized houses that were purchased at different times. What can you do to help decide whether to pay the $230,000?

Since you're taking an econometrics class, you decide to collect data on all local houses that were sold within the last few weeks and to build a regression model of the sales prices of the houses as a function of their sizes.[10] Such a data set is called **cross-sectional** because all of the observations are from the same point in time and represent different individual economic entities (like countries or, in this case, houses) from that same point in time.

To measure the impact of size on price, you include the size of the house as an independent variable in a regression equation that has the price of that house as the dependent variable. You expect a positive sign for the coefficient of size, since big houses cost more to build and tend to be more desirable than small ones. Thus the theoretical model is:

$$\text{PRICE}_i = \beta_0 + \underset{+}{\beta_1}\text{SIZE}_i + \epsilon_i \tag{1.20}$$

where: PRICE_i = the price (in thousands of $) of the *i*th house
SIZE_i = the size (in square feet) of that house
ϵ_i = the value of the stochastic error term for that house

You collect the records of all recent real estate transactions, find that 43 local houses were sold within the last 4 weeks, and estimate the following regression of those 43 observations:

$$\widehat{\text{PRICE}}_i = 40.0 + 0.138\text{SIZE}_i \tag{1.21}$$

What do these estimated coefficients mean? The most important coefficient is $\hat{\beta}_1 = 0.138$, since the reason for the regression is to find out the impact of size on price. This coefficient means that if size increases by 1 square foot,

10. It's unusual for an economist to build a model of price without including some measure of quantity on the right-hand side. Such models of the price of a good as a function of the attributes of that good are called *hedonic* models and will be discussed in greater depth in Section 11.8. The interested reader is encouraged to skim the first few paragraphs of that section before continuing on with this example.

price will increase by 0.138 thousand dollars ($138). $\hat{\beta}_1$ thus measures the change in $PRICE_i$ associated with a one-unit increase in $SIZE_i$. It's the slope of the regression line in a graph like Figure 1.5.

What does $\hat{\beta}_0 = 40.0$ mean? $\hat{\beta}_0$ is the estimate of the constant or intercept term. In our equation, it means that price equals 40.0 when size equals zero. As can be seen in Figure 1.5, the estimated regression line intersects the price axis at 40.0. While it might be tempting to say that the average price of a vacant lot is $40,000, such a conclusion would be unjustified for a number of reasons, which will be discussed in Section 7.1. It's much safer either to interpret $\hat{\beta}_0 = 40.0$ as nothing more than the value of the estimated regression when $S_i = 0$, or to not interpret $\hat{\beta}_0$ at all.

What does $\hat{\beta}_1 = 0.138$ mean? $\hat{\beta}_1$ is the estimate of the coefficient of SIZE in Equation 1.20, and as such it's also an estimate of the slope of the line in Figure 1.5. It implies that an increase in the size of a house by one square foot will cause the estimated price of the house to go up by 0.138 thousand dollars or $138. It's a good habit to analyze estimated slope coefficients to see whether they make sense. The positive sign of $\hat{\beta}_1$ certainly is what we expected, but what about the magnitude of the coefficient? Whenever you interpret a coefficient, be sure to take the units of measurement into consideration. In this case, is $138 per square foot a plausible number? Well, it's

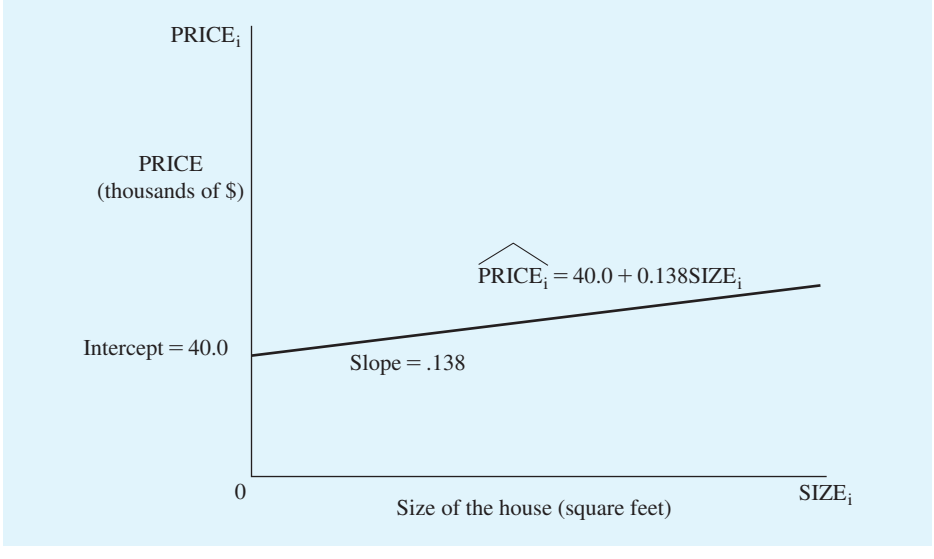


Figure 1.5 A Cross-Sectional Model of Housing Prices

A regression equation that has the price of a house as a function of the size of that house has an intercept of 40.0 and a slope of 0.138, using Equation 1.21.

hard to know for sure, but it certainly is a lot more reasonable than $1.38 per square foot or $13,800 per square foot!

How can you use this estimated regression to help decide whether to pay $230,000 for the house? If you calculate a $\hat{Y}$ (predicted price) for a house that is the same size (1,600 square feet) as the one you're thinking of buying, you can then compare this $\hat{Y}$ with the asking price of $230,000. To do this, substitute 1600 for $SIZE_i$ in Equation 1.21, obtaining:

$$\widehat{PRICE}_i = 40.0 + 0.138(1600) = 40.0 + 220.8 = 260.8$$

The house seems to be a good deal. The owner is asking "only" $230,000 for a house when the size implies a price of $260,800! Perhaps your original feeling that the price was too high was a reaction to steep housing prices in general and not a reflection of this specific price.

On the other hand, perhaps the price of a house is influenced by more than just the size of the house. Such multivariate models are the heart of econometrics, and we'll add more independent variables to Equation 1.21 when we return to this housing price example in Section 11.8.

1.6 Summary

1. Econometrics—literally, "economic measurement"—is a branch of economics that attempts to quantify theoretical relationships. Regression analysis is only one of the techniques used in econometrics, but it is by far the most frequently used.
2. The major uses of econometrics are description, hypothesis testing, and forecasting. The specific econometric techniques employed may vary depending on the use of the research.
3. While regression analysis specifies that a dependent variable is a function of one or more independent variables, regression analysis alone cannot prove or even imply causality.
4. A stochastic error term must be added to all regression equations to account for variations in the dependent variable that are not explained completely by the independent variables. The components of this error term include:
   a. omitted or left-out variables
   b. measurement errors in the data
   c. an underlying theoretical equation that has a different functional form (shape) than the regression equation
   d. purely random and unpredictable events

5. An estimated regression equation is an approximation of the true equation that is obtained by using data from a sample of actual Ys and Xs. Since we can never know the true equation, econometric analysis focuses on this estimated regression equation and the estimates of the regression coefficients. The difference between a particular observation of the dependent variable and the value estimated from the regression equation is called the residual.

EXERCISES

(The answers to the even-numbered exercises are in Appendix A.)

1. Write the meaning of each of the following terms without referring to the book (or your notes), and compare your definition with the version in the text for each:
   a. constant or intercept (p. 7)
   b. cross-sectional (p. 21)
   c. dependent variable (p. 5)
   d. estimated regression equation (p. 14)
   e. expected value (p. 9)
   f. independent (or explanatory) variable (p. 5)
   g. linear (p. 8)
   h. multivariate regression model (p. 12)
   i. regression analysis (p. 5)
   j. residual (p. 15)
   k. slope coefficient (p. 7)
   l. stochastic error term (p. 8)

2. Use your own computer's regression software and the weight (Y) and height (X) data from Table 1.1 to see if you can reproduce the estimates in Equation 1.19. There are two ways to load the data: You can type in the data yourself or you can download datafile HTWT1 (in Stata, EViews, Excel, or ASCII formats) from the text's website: http://www.pearsonhighered.com/studenmund. Once the datafile is loaded, run $Y = f(X)$, and your results should match Equation 1.19. Different programs require different commands to run a regression. For help in how to do this with Stata or EViews, either see the answer to this question in Appendix A or read Appendix 1.7.

3. Not all regression coefficients have positive expected signs. For example, a *Sports Illustrated* article by Jaime Diaz reported on a study of golfing putts of various lengths on the Professional Golfers' Association (PGA) Tour.[11] The article included data on the percentage of putts made (P_i) as a function of the length of the putt in feet (L_i). Since the longer the putt, the less likely even a professional is to make it, we'd expect L_i to have a negative coefficient in an equation explaining P_i. Sure enough, if you estimate an equation on the data in the article, you obtain:

$$\hat{P}_i = 83.6 - 4.1L_i \qquad (1.22)$$

   a. Carefully write out the exact meaning of the coefficient of L_i.
   b. Suppose someone else took the data from the article and estimated:

$$P_i = 83.6 - 4.1L_i + e_i$$

      Is this the same result as that of Equation 1.22? If so, what definition do you need to use to convert this equation back to Equation 1.22?
   c. Use Equation 1.22 to determine the percent of the time you'd expect a PGA golfer to make a 10-foot putt. Does this seem realistic? How about a 1-foot putt or a 25-foot putt? Do these seem as realistic?
   d. Your answer to part c should suggest that there's a problem in applying a linear regression to these data. What is that problem?

4. Return to the housing price model of Section 1.5 and consider the following equation:

$$\widehat{SIZE}_i = -290 + 3.62\ PRICE_i \qquad (1.23)$$

   where: $SIZE_i$ = the size (in square feet) of the *i*th house
   $PRICE_i$ = the price (in thousands of $) of that house

   a. Carefully explain the meaning of each of the estimated regression coefficients.
   b. Suppose you're told that this equation explains a significant portion (more than 80 percent) of the variation in the size of a house. Have we shown that high housing prices cause houses to be large? If not, what have we shown?
   c. What do you think would happen to the estimated coefficients of this equation if we had measured the price variable in dollars instead of in thousands of dollars? Be specific.

11. Jaime Diaz, "Perils of Putting," *Sports Illustrated,* April 3, 1989, pp. 76–79.

5. If an equation has more than one independent variable, we have to be careful when we interpret the regression coefficients of that equation. Think, for example, about how you might build an equation to explain the amount of money that different states spend per pupil on public education. The more income a state has, the more they probably spend on public schools, but the faster enrollment is growing, the less there would be to spend on each pupil. Thus, a reasonable equation for per pupil spending would include at least two variables: income and enrollment growth:

$$S_i = \beta_0 + \beta_1 Y_i + \beta_2 G_i + \epsilon_i \qquad (1.24)$$

where: S_i = educational dollars spent per public school student in the *i*th state
Y_i = per capita income in the *i*th state (in dollars)
G_i = the percent growth of public school enrollment in the *i*th state

a. State the economic meaning of the coefficients of Y and G. (*Hint:* Remember to hold the impact of the other variable constant.)
b. If we were to estimate Equation 1.24, what signs would you expect the coefficients of Y and G to have? Why?
c. Silva and Sonstelie estimated a cross-sectional model of per student spending by state that is very similar to Equation 1.24:[12]

$$\hat{S}_i = -183 + 0.1422 Y_i - 5926 G_i \qquad (1.25)$$
$$N = 49$$

Do these estimated coefficients correspond to your expectations? Explain Equation 1.25 in common sense terms.
d. The authors measured G as a decimal, so if a state had a 10 percent growth in enrollment, then G equaled .10. What would Equation 1.25 have looked like if the authors had measured G in percentage points, so that if a state had 10 percent growth, then G would have equaled 10? (*Hint:* Write out the actual numbers for the estimated coefficients.)

6. Your friend has an on-campus job making telephone calls to alumni asking for donations to your college's annual fund, and she wonders

12. Fabio Silva and Jon Sonstelie, "Did Serrano Cause a Decline in School Spending?" *National Tax Review,* Vol. 48, No. 2, pp. 199–215. The authors also included the tax price for spending per pupil in the *i*th state as a variable.

whether her calling is making any difference. In an attempt to measure the impact of student calls on fund raising, she collects data from 50 alums and estimates the following equation:

$$\widehat{GIFT}_i = 2.29 + 0.001INCOME_i + 4.62CALLS_i \qquad (1.26)$$

where:
- $GIFT_i$ = the 2016 annual fund donation (in dollars) from the *i*th alum
- $INCOME_i$ = the 2016 estimated income (in dollars) of the *i*th alum
- $CALLS_i$ = the # of calls to the *i*th alum asking for a donation in 2016

a. Carefully explain the meaning of each estimated coefficient. Are the estimated signs what you expected?
b. Why is the left-hand variable in your friend's equation $\widehat{GIFT}_i$ and not $GIFT_i$?
c. Your friend didn't include the stochastic error term in the estimated equation. Was this a mistake? Why or why not?
d. Suppose that your friend decides to change the units of INCOME from "dollars" to "thousands of dollars." What will happen to the estimated coefficients of the equation? Be specific.
e. If you could add one more variable to this equation, what would it be? Explain.

7. Let's return to the wage determination example of Section 1.2. In that example, we built a model of the wage of the *i*th worker in a particular field as a function of the work experience, education, and gender of that worker:

$$WAGE_i = \beta_0 + \beta_1 EXP_i + \beta_2 EDU_i + \beta_3 GEND_i + \epsilon_i \qquad (1.10)$$

where:
- $Y_i = WAGE_i$ = the wage of the *i*th worker
- $X_{1i} = EXP_i$ = the years of work experience of the *i*th worker
- $X_{2i} = EDU_i$ = the years of education beyond high school of the *i*th worker
- $X_{3i} = GEND_i$ = the gender of the *i*th worker (1 = male and 0 = female)

a. What is the real-world meaning of β_2? (*Hint:* If you're unsure where to start, review Section 1.2.)
b. What is the real-world meaning of β_3? (*Hint:* Remember that GEND is a dummy variable.)

c. Suppose that you wanted to add a variable to this equation to measure whether there might be discrimination against people of color. How would you define such a variable? Be specific.
d. Suppose that you had the opportunity to add another variable to the equation. Which of the following possibilities would seem best? Explain your answer.

   i. the age of the *i*th worker
   ii. the number of jobs in this field
   iii. the average wage in this field
   iv. the number of "employee of the month" awards won by the *i*th worker
   v. the number of children of the *i*th worker

8. Have you heard of "RateMyProfessors.com"? On this website, students evaluate a professor's overall teaching ability and a variety of other attributes. The website then summarizes these student-submitted ratings for the benefit of any student considering taking a class from the professor.

   Two of the most interesting attributes that the website tracks are how "easy" the professor is (in terms of workload and grading), and how "hot" the professor is (presumably in terms of physical attractiveness). An article by Otto and colleagues[13] indicates that being "hot" improves a professor's rating more than being "easy." To investigate these ideas ourselves, we created the following equation for RateMyProfessors.com:

$$\text{RATING}_i = \beta_0 + \beta_1 \text{EASE}_i + \beta_2 \text{HOT}_i + \epsilon_i \qquad (1.27)$$

   where: RATING_i = the overall rating (5 = best) of the *i*th professor
   EASE_i = the easiness rating (5 = easiest) of the *i*th professor
   HOT_i = 1 if the *i*th professor is considered "hot," 0 otherwise

   To estimate Equation 1.27, we need data, and Table 1.2 contains data for these variables from 25 randomly chosen professors on RateMyProfessors.com. If we estimate Equation 1.27 with the data in Table 1.2, we obtain:

$$\widehat{\text{RATING}}_i = 3.23 + 0.01\text{EASE}_i + 0.59\text{HOT}_i \qquad (1.28)$$

13. James Otto, Douglas Sanford, and Douglas Ross, "Does RateMyProfessors.com Really Rate My Professor?" *Assessment and Evaluation in Higher Education*, August 2008, pp. 355–368.

Table 1.2 RateMyProfessors.com Ratings

Observation	RATING	EASE	HOT
1	2.8	3.7	0
2	4.3	4.1	1
3	4.0	2.8	1
4	3.0	3.0	0
5	4.3	2.4	0
6	2.7	2.7	0
7	3.0	3.3	0
8	3.7	2.7	0
9	3.9	3.0	1
10	2.7	3.2	0
11	4.2	1.9	1
12	1.9	4.8	0
13	3.5	2.4	1
14	2.1	2.5	0
15	2.0	2.7	1
16	3.8	1.6	0
17	4.1	2.4	0
18	5.0	3.1	1
19	1.2	1.6	0
20	3.7	3.1	0
21	3.6	3.0	0
22	3.3	2.1	0
23	3.2	2.5	0
24	4.8	3.3	0
25	4.6	3.0	0

Datafile = RATE1

a. Take a look at Equation 1.28. Do the estimated coefficients support our expectations? Explain.

b. See if you can reproduce the results in Equation 1.28 on your own. To do this, take the data in Table 1.2 and use Stata or your own regression program to estimate the coefficients from these data. If you do everything correctly, you should be able to verify the estimates in Equation 1.28. (If you're not sure how to get started on this question, either take a look at the answer to Exercise 2 in Appendix A or read Appendix 1.7.)

c. This model includes two independent variables. Does it make sense to think that the teaching rating of a professor depends on just these two variables? What other variable(s) do you think might be important?
d. Suppose that you were able to add your suggested variable(s) to Equation 1.28. What do you think would happen to the coefficients of EASE and HOT when you added the variable(s)? Would you expect them to change? Would you expect them to remain the same? Explain.
e. (optional) Go to the RateMyProfessors.com website, choose 25 observations at random, and estimate your own version of Equation 1.27. Now compare your regression results to those in Equation 1.28. Do your estimated coefficients have the same signs as those in Equation 1.28? Are your estimated coefficients exactly the same as those in Equation 1.28? Why or why not?

1.7 Appendix: Using Stata

Using Econometrics is about, well, using econometrics, and it doesn't take long to realize that using econometrics requires software. The powerful and user-friendly econometric software package referred to in the text is Stata[14], and the purpose of this appendix[15] is to give you a brief introduction to Stata.

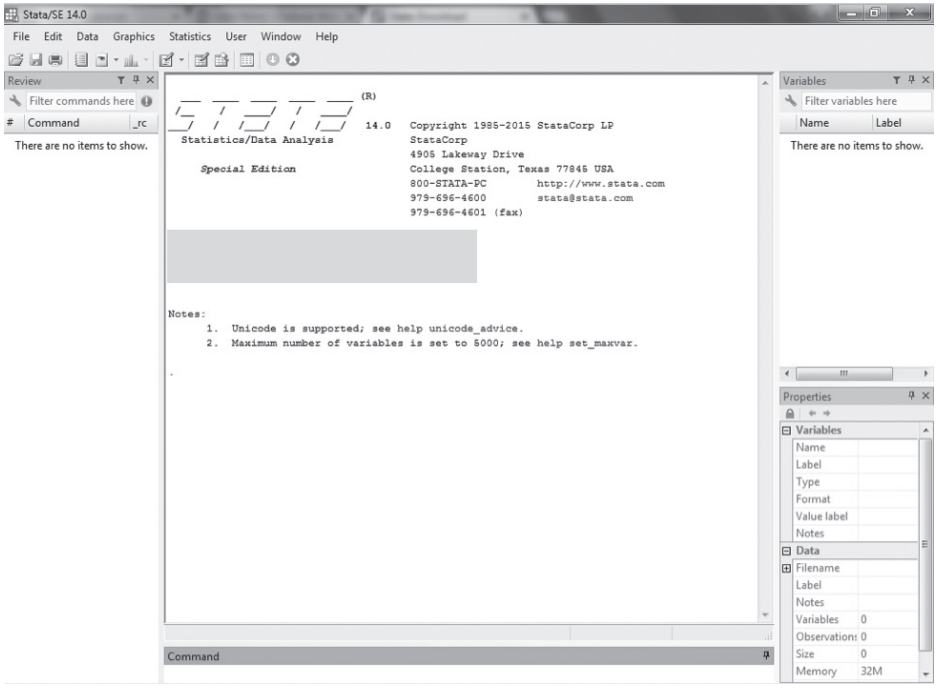
For most people (including me!), learning new computer software involves some pain. Our goal in this Appendix is to take away as much of that pain as possible. We hope to give you a head start with Stata and also convince you that it's worth your time to check out the complete "Using Stata" document found online at the *Using Econometrics* student companion website (http://www.pearsonhighered.com/studenmund). That free document (yes, free!) is designed to get you up and running in Stata with as little pain as possible. It shows in plain English and clear pictures how to use all the econometric techniques you'll encounter in the text (and more!)

How do you get Stata? There are a number of ways. Your college or university may provide Stata access in official computer labs. If it doesn't (or if you want a personal copy), you can buy and download Stata directly (http://www.stata.com). Fortunately, reasonable student pricing is available.

14. Other econometric software programs that you might encounter include EViews, SAS, R, and SPSS.

15. Written by John Perry, Centre College. Used with permission.

With access to Stata, go ahead and "open" it as you would any program on your computer (like Word, Excel, etc.). When you open Stata on a PC, you should see something like this:



Stata also runs on a Mac, and while it looks slightly different, the commands and functionality are almost the same as on a PC.

Let's talk about what you see. There are five "windows" within Stata. The biggest one, squarely in the middle of the screen, is the "Results" window. Nicely, it shows you the results of what you tell Stata to do.

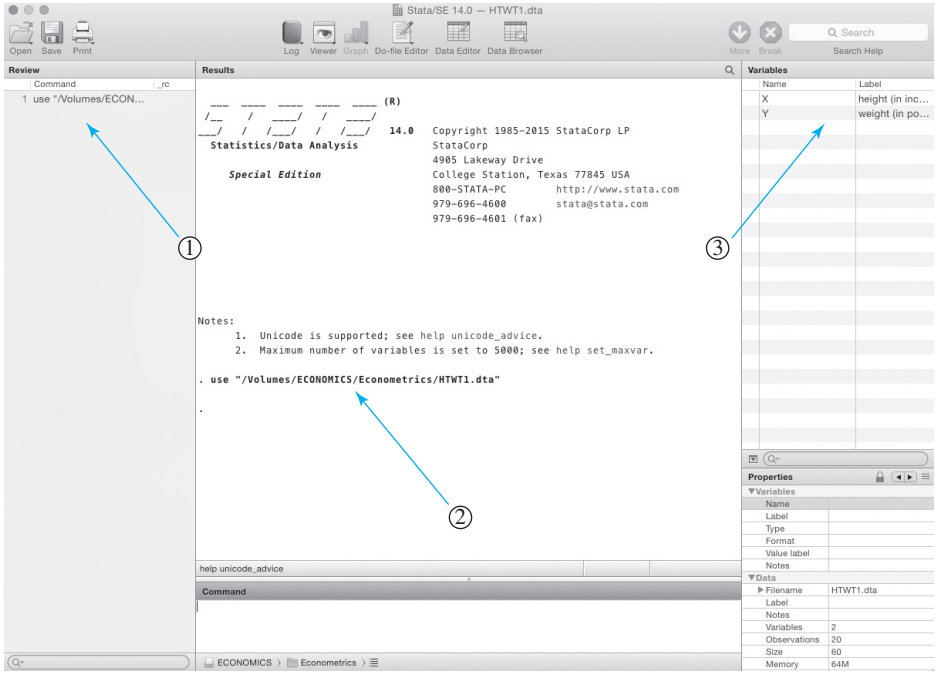
At the top left is the "Review" window. This window shows a history of all the commands you have given Stata. The top right is where the variables in your dataset will show up and the bottom right is where you'll see properties of the variables.

The bottom, center window is the "Command" window. As the name suggests, this is where you tell Stata what to do, where you actually "program." (Don't panic! You can work in Stata by typing commands one at a time or you can roll all your comments up into a single program—called in Stata language a "do-file." The full "Using Stata" document covers do-files.)

With Stata open, we should move along and open a dataset. In Section 1.4, you met a dataset from Magic Hill amusement park named HTWT1.dta (".dta" is the format of a Stata dataset much like "docx" is the format for a Microsoft Word document). It contained the height and weight of 20 people where:

Y_i = weight (in pounds) of the *i*th customer
X_i = height (in inches above 5 feet) of the *i*th customer

You can (and should at this point) download and save the dataset to your computer from the student companion website. After doing that, to open the dataset, go to the top left in Stata and click on the folder icon. Next, you'll be guided to find where you saved HTWT1.dta. Highlight it and click "open." This is similar to how you'd open a file in any other software (like Word, Excel, etc.). You should see something like this (this time we used Stata on a Mac):



Notice that what you commanded Stata to do—to open HTWT1.dta—is recorded in the Results window (indicated by arrow 1). In Stata, "use" means open. The "use" statement is followed by the pathname (in quotes) where the file is saved on your computer (in my case "/Volumes/ECONOMICS/

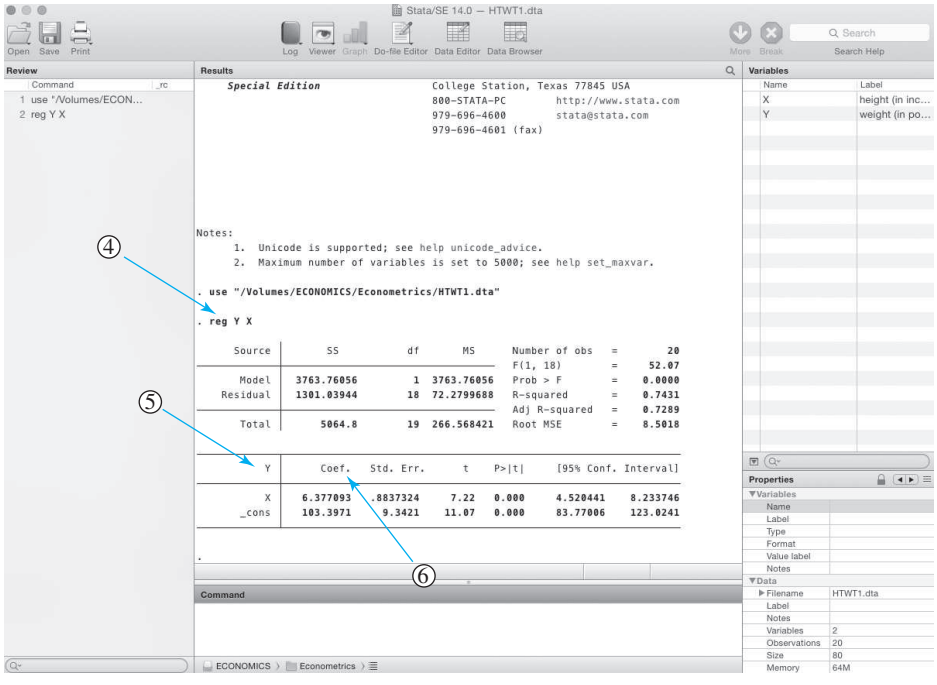
Econometrics/HTWT1.dta"). This command is also recorded in the Review window and indicated by arrow 2.

At the top right, signaled by arrow 3, you see that you have two variables in your Variables window (X and Y). This means you now have data in Stata.

Things are about to get exciting! With our data open in Stata we're now in a position to replicate Equation 1.19. To do so, type "reg Y X" into the Command window in Stata and hit enter.

The "reg" command, which is short for "regress," tells Stata to perform a regression. Directly after "reg," insert the dependent variable (Y in our case). The dependent variable is followed by the model's independent variables. Equation 1.19 has one independent variable named X. Note that Stata is case sensitive. If you type "y" when the variable's name is "Y," Stata will yell.

After giving the "reg Y X" command, you should see something like this:



What you see in the Results window above could easily overwhelm a person. For now, focus on where the three arrows direct. Arrow 4 points to the command that had Stata produce the estimation. Arrow 5 points to the column that lists the variables in the regression: Y, X, and something called "_cons". That "something" is the model's intercept term, otherwise known as β_0.

Arrow 6 points to the "Coef." column, which reports the estimated coefficients. The first number in the Coef. column is 6.377093. That is $\hat{\beta}_1$, the coefficient estimate for X, and matches the 6.38 (rounded) of Equation 1.19. Moving down the Coef. column and next to the _cons is 103.3971. That is $\hat{\beta}_0$, the estimate of the intercept, which rounds to 103.40.

And with that, you've estimated your first regression in Stata! Keep in mind, however, that this short appendix is meant only to help get you started in Stata. The full "Using Stata" document will show you much more—while trying to minimize the pain.